

CLAIMS

1. A method of executing a query on a data repository, the method comprising:
receiving a query for execution on data in the data repository;
generating an estimate of a number of results of the query;
defining a subset of data in the data repository;
determining whether to execute the query on the subset of the data;
if the query is to be executed on the subset of the data, executing the query on the subset of the data to generate a partial set of results, otherwise executing the query on the data repository to generate a complete set of results; and
providing query results.
2. A method in accordance with claim 1, wherein providing query results comprises making the query results available to an application program.
3. A method in accordance with claim 2, further comprising:
the application program providing query results to a user interface.
4. A method in accordance with claim 1, wherein determining whether to execute the query on the subset of the data comprises determining whether a sufficient number of results will be generated by executing the query on the subset of the data.
5. A method in accordance with claim 1, wherein determining whether to execute the query on the subset of the data comprises estimating whether executing the query on the subset of the data would generate a desired number of results, the method further comprising:
receiving a value representing the desired number of results.
6. A method in accordance with claim 1, wherein:
the method further comprises receiving a value representing the desired number of results;

the query is to be executed on the subset of the data if the estimate of the number of results of the query is greater than a weighted subset estimate generated in accordance with the following estimation function:

$$R * \frac{N}{stripeSize} * F, \text{ where } R \text{ is the number of results desired, } N \text{ is the total number of}$$

possible results, F is an arbitrary number, and $stripeSize$ is the size of the subset of the data; and

determining whether to execute the query on the subset of the data comprises:

generating the weighted subset estimate; and

determining whether the estimate of the number of results of the query is greater than the weighted subset estimate.

7. A method in accordance with claim 1 further comprising:

in response to executing the query on an $(N - 1)$ th subset of the data, determining whether a sufficient number of results have been generated; and

if a sufficient number of results have been generated, defining an N th subset of the data in the data repository and executing the query on the N th subset of the data, otherwise executing the query on the data repository.

8. A method in accordance with claim 1, wherein generating an estimate of a number of results of the query is generated in accordance with the following estimation functions:

$$est(NOT) = N - est(op),$$

$$est(AND) = \frac{est(op_1) * est(op_2)}{N}, \text{ and}$$

$$est(OR) = est(op_1) + est(op_2) - \frac{est(op_1) * est(op_2)}{N},$$

where op is an operand, $est()$ signifies an estimate of the operator or operand in the parenthesis, and N is the total number of possible results.

9. An information management system, the system comprising:
a data repository, wherein the data repository is configured to store data; and
one or more processes for executing queries on the data repository, wherein the one or more processes are operative to:
receive a query for execution on data in the data repository;
generate an estimate of a number of results of the query;
define a subset of data in the data repository;
determine whether to execute the query on the subset of the data;
if the query is to be executed on the subset of the data, execute the query on the subset of the data to generate a partial set of results, otherwise execute the query on the data repository to generate a complete set of results; and
provide query results.
10. An information management system in accordance with claim 9, wherein the operation of determining whether to execute the query on the subset of the data comprises determining whether a sufficient number of results will be generated by executing the query on the subset of the data.
11. An information management system in accordance with claim 9, wherein the operation of providing query results comprises making the query results available to an application program.
12. An information management system in accordance with claim 9, wherein the operation of determining whether to execute the query on the subset of the data comprises estimating whether executing the query on the subset of the data would generate a desired number of results, the one or more processes are further operative to:
receive a value representing the desired number of results.
13. An information management system in accordance with claim 9, wherein the one or more processes are further operative to:

in response to executing the query on an (N - 1)th subset of the data, determine whether a sufficient number of results have been generated; and

if a sufficient number of results have been generated, define an Nth subset of the data in the data repository and execute the query on the Nth subset of the data, otherwise execute the query on the data repository.

14. A computer program product, tangibly embodied on an information carrier, the computer program product comprising instructions operable to cause data processing apparatus to:

receive a query for execution on data in a data repository;

generate an estimate of a number of results of the query;

define a subset of data in the data repository;

determine whether to execute the query on the subset of the data;

if the query is to be executed on the subset of the data, execute the query on the subset of the data to generate a partial set of results, otherwise execute the query on the data repository to generate a complete set of results; and

provide query results.

15. A computer program product in accordance with claim 14, wherein the operation of providing query results comprises making the query results available to an application program.

16. A computer program product in accordance with claim 14, wherein the operation of determining whether to execute the query on the subset of the data comprises determining whether a sufficient number of results will be generated by executing the query on the subset of the data.

17. A computer program product in accordance with claim 14, wherein the operation of determining whether to execute the query on the subset of the data comprises estimating whether executing the query on the subset of the data would generate a desired number of results, the computer program product further comprising instructions operable to:

receive a value representing the desired number of results.

18. A computer program product in accordance with claim 14, wherein the computer program product further comprises instructions operable to:

in response to executing the query on an $(N - 1)$ th subset of the data, determine whether a sufficient number of results have been generated; and

if a sufficient number of results have been generated, define an Nth subset of the data in the data repository and execute the query on the Nth subset of the data, otherwise execute the query on the data repository.